CLASSiC

D6.5 Annotated Data Archive

Olivier Pietquin, Helen Hastie, Srini Janarthanam,
Simon Keizer, Ghislain Putois, Lonneke van der Plas

Distribution: Public

CLASSiC
Computational Learning in Adaptive Systems for Spoken Conversation
216594 Deliverable 6.5

April 2011

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page.

Project ref. no.	216594
Project acronym	CLASSiC
Project full title	Computational Learning in Adaptive Systems for Spoken Conversation
Instrument	STREP
Thematic Priority	Cognitive Systems, Interaction, and Robotics
Start date / duration	01 March 2008 / 36 Months

Security	Public
Contractual date of delivery	M36 = February 2011
Actual date of delivery	April 2011
Deliverable number	6.5
Deliverable title	D6.5 Annotated Data Archive
Type	Report
Status & version	Final 1.0
Number of pages	12 (excluding front matter)
Contributing WP	3
WP/Task responsible	SUPELEC
Other contributors	ALL
Author(s)	Olivier Pietquin, Helen Hastie, Srini Janarthanam, Simon Keizer, Ghislain Putois, Lonneke van der Plas
EC Project Officer	Philippe Gelin
Keywords	Data, Corpora, Spoken Dialogue

The partners in CLASSiC are:

Heriot-Watt University	HWU
University of Cambridge	UCAM
University of Geneva	GENE
Ecole Superieure d'Electricite	SUPELEC
France Telecom/ Orange Labs	FT
University of Edinburgh HCRC	EDIN

For copies of reports, updates on project activities and other CLASSIC-related information, contact:

The CLASSIC Project Co-ordinator:
Dr. Oliver Lemon
School of Mathematical and Computer Sciences (MACS)
Heriot-Watt University
Edinburgh
EH14 4AS
United Kingdom
O.Lemon@hw.ac.uk
Phone +44 (131) 451 3782 - Fax +44 (0)131 451 3327

Copies of reports and other material can also be accessed via the project's administration homepage, `http://www.classic-project.org`

©2011, The Individual Authors.

No part of this document may be reproduced or transmitted in any form, or by any means, electronic

Contents

Executive summary

This document presents the work done in the framework of the CLASSiC project task 6.5 and the results of the associate deliverable 6.5. In particular, this deliverable concerns the data collected by the different partners during the project. It gives a summary of the different data sets that have been made available to the public for free download at the website `http://www.macs.hw.ac.uk/ilabarchive/classicproject/data/`

Chapter 1

Introduction

The aim of the CLASSiC project is the statistical optimization of spoken dialogue systems. Therefore, different databases have been developed during the project, each with different goals. First, statistical optimization requires training algorithms on actual data. Second, evaluation of performance requires releasing the developed systems to users and collecting information about the interactions.

Also, since every sub-system had to be optimized, in addition to the whole dialogue systems, several types of data were collected, for example Spoken Language Understanding, Dialogue Management, and Natural Language Generation.

The collected data are freely and publicly released, which constitutes an important oucome of the project. The CLASSiC project data is hosted in a web-based repository at Heriot-Watt University. The URL is:

`http://www.macs.hw.ac.uk/ilabarchive/classicproject/data/`

To be able to download any data, a user is required to first register with the system by providing a valid email address and a password chosen by the user. An activation email will then be sent to the user's email address which contains the activation link. By clicking the activation link the user account will be activated, and then the user can log on to the system to download files.

Chapter 2

Datasets Description

2.1 EuroParl

Partner in charge : UNIGE

Description : The EuroParl [3] parallel corpus is the proceedings of the European Parliament (English and French texts)[1]. These are transcribed debates, however, the spoken language is not spontaneous but rather controlled.

We have manually annotated 1000 French sentences with semantic roles and predicates using the annotation scheme of PropBank and the format used in the CoNLL-2008 shared task. Our syntactic parser was used to provide the syntactic structure of these sentences. We automatically annotated the English and French parallel sentences from the EuroParl corpus with syntactic and semantic annotations using statistical syntactic-semantic parsers developed for English and French.

Type : Text

Recording conditions : N/A

Experimental setup : N/A

Type of annotation : Manual annotations of semantic roles and predicates, automatic annotation of syntactic dependencies on 1000 French sentences. The remainder (300k up to 2 million sentences -depending on whether we discard indirect translations- of English and French) is annotated automatically.

Annotation scheme : Format and annotation scheme from the CoNLL 2008 shared task[2] [11] with the exception of the NomBank annotations. We do not annotate nominal predicates (NomBank), just verbal predicates as in PropBank [7].

Size : 1MB of manually annotated data and 500MB up to 3GB of automatically annotated data (depending on whether we discard indirect translations)

Original aim of data : Training and evaluating syntactic-semantic parsers

[1] http://www.statmt.org/europarl/archives.html

[2] http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=conll2008:format

2.2 TownInfo Evaluation Dialogues - HIS

Partner in charge : UCAM

Description : Town info domain dialogue system (tourist information about a fictitious town). Dialogues including ASR audio, dialogue log files and transcriptions.

Type : Audio and text

Recording conditions : Desktop microphone

Experimental setup : Real system (HIS). Subjects were recruited to come to the lab and talk to an end-to-end spoken dialogue system on a desktop computer using a head set; each subject performed a number of tasks following predefined scenarios.

Type of annotation : Manual audio transcription and semantic annotation.

Annotation scheme : Recruited British native speakers; semantic annotation using Cambridge dialogue act specification and an in-lab instruction session.

Size : 2.2GB (720 dialogues)

Original aim of data : Dialogue system evaluation

Date of collection : March 2009

2.3 CamInfo Evaluation Dialogues - HIS

Partner in charge : UCAM

Description : Cambridge info domain dialogue system (information about places to eat in Cambridge). Dialogues including ASR audio, dialogue log files and transcriptions

Type : Audio and text

Recording conditions : Telephone or Skype

Experimental setup : Real system. Subjects were recruited to talk to an end-to-end spoken dialogue system using a telephone connection (mobile, landline, or Skype); each subject performed a number of tasks following predefined scenarios

Type of annotation : Manual audio transcription using Amazon Mechanical Turk; semantic transcription obtained by automatic semantic parsing of audio transcriptions.

Annotation scheme : MTurk

Size : 3.9GB (709 dialogues); Feb'11 corpus: system evaluation 6.0G (1338 dialogues)

Original aim of data : Dialogue system evaluation

Date of collection : November 2010 and February 2011

2.4 TownInfo - DIPPER

Partner in charge : UEDIN

Description : Town info domain dialogue system (tourist information about a fictitious town). Dialogues including ASR audio, dialogue log files. Fluent English speakers. 23 Subjects.

Type : Audio and text.

Recording conditions : Headset microphone

Experimental setup : Real system (DIPPER). Subjects were recruited to come to the lab and talk to an end-to-end spoken dialogue system on a desktop computer using a head set; each subject performed a number of tasks following predefined scenarios.

Type of annotation : Logs contain automatic semantic annotation.

Annotation scheme : Automatic semantic annotation of user acts using University of Edinburgh (TALK) dialogue act specification

Size : 120MB (275 dialogues)

Original aim of data : Dialogue systems evaluation; comparison of DIPPER-MDP, DIPPER-QMDP and DIPPER-POMDP systems

Date of collection : May-June 2009

2.5 Appointment Scheduling System 2 evaluation dialogues

Partner in charge : UCAM

Description : French appointment scheduling data. Dialogues including ASR audio, dialogue log files and transcriptions

Type : Audio and text

Recording conditions : Telephone

Experimental setup : Real system.

Type of annotation : Manual audio transcription.

Annotation scheme : Recruit native speakers to transcribe audio

Size : 1.1GB.

Original aim of data : CLASSiC system 2 data collection and evaluation

Date of collection : November 2010 and February 2011

2.6 Temporal Referring Expressions

Partner in charge : HWU

Description : A web-based experiment was conducted to gather user feedback and understanding of a variety of Temporal Expressions (TE). The data collection experiment was in 2 parts (Task 1 and Task 2) and was designed using the Webexp experimental software[3]. Webexp is a client-server set up where a server application hosts the experiment and stores the experimental files, logs and results. The client side runs an applet on the user's web-browser. Participants who completed both tasks were rewarded by a chance to win one of three £50 Amazon vouchers.

In Task 1, participants listened to an audio file containing a TE generated from absolute and relative TE units (TEUs). No relative-context TEUs were used in Task 1 since the dialogue excerpt presented was in isolation and had no context. Each participant was asked to listen to 10 different audio files for 10 different dates for which the TE was randomly picked from a set of 30 combinations of TEUs. Each TE was generated by a rule-based realiser and synthesized using Baratinoo synthesizer. This realiser generates text from a candidate list for each TEU based on a given date. For example, if the slot currently being discussed is Tuesday 7th in the pm, the realiser would generate "tomorrow" for DAY=rel but "the day after tomorrow" for Wednesday 8th am. The participant then had to identify the correct slot that the system is referring to in a 2-week calendar.

Task 2 of the experiment was in two stages. In the first stage (Task 2A), the participants are given today's date and the following dialogue excerpt "Operator: We need to send out an engineer to your home. The first available appointment is:". They are then asked to listen to 5 audio files of the system saying different TEs for the same date and asked to rate preference on a scale of 1-6 (where 1 is bad and 6 is great.) For the second stage (Task 2B), the dialogue is as follows "Operator: so you can't do Wednesday 8th September in the morning" and then the participants are asked to listen to 6 more audio files that are generated TEs including relative context such as "how about Thursday at the same time?". This two-stage process is then repeated 4 times for each participant.

For each Task there are the following directories/files:

- logs: webexp log files e.g. time stamped audio play,
- results: XML files of participant time stamped activity, e.g. preference ratings,
- subjects: information about the participant in XML, e.g. browser used,
- overview: csv overview of statistics gathered.

Type : Log files of a webexp on-line experiments.

Recording conditions : N/A

Experimental setup : Webexp on-line experiment.

Type of annotation : Automatic.

[3] http www.webexp.info

Annotation scheme : TEs are broken down into 5 categories or units (TEUs): day (DY), date (DD), month (MM), week (WK) and time (TM). Each of these units can be defined in terms relative to the current day and to the current context (i.e. previously mentioned dates). Specifically, there are 3 unit attributes: absolute (e.g. DAY=abs "Tuesday"); relative to current day (e.g. DAY=rel "tomorrow"); and relative to context (e.g. DAY=rc "the following day").

Size : 20 MB. In total there were 73 participants for Task 1 and 730 TE samples collected. Although Task 2 directly followed on from Task 1, there was a significant drop out rate as only 48 participants completed the second task resulting in 1920 TE samples.

Original aim of data : The aim of the web-based experiment was to gather data on user understanding and preferences for Temporal Expressions (TE) of varying types in different contexts. Metrics such as user preference and apparent ambiguity were collected and used to build a user simulator for an Appointment Scheduling spoken dialogue system. A policy for optimal TE generation was trained using this user simulator. This policy was evaluated in simulation and with real users and shows significant improvement over hand-coded policies, see [2].

Date of collection : September 2010.

2.7 Self-Help dialogues

Partner in charge : UEDIN

Description : Dialogue between users and Wizard of Oz system on the topic of setting broadband Internet connection.

- Training Corpus: Users with different levels of domain knowledge in setting up broadband Internet connections interacted with a non-adaptive wizarded dialogue system that used three different types of NLG policies. The system gave users instructions to set up the Internet connection using the equipments and objects in front of the users.
- Evaluation Corpus: Users with different levels of domain knowledge in setting up broadband Internet connections interacted with two adaptive wizarded dialogue systems. One of them used a hand-coded NLG policy and the other used a learned adaptive NLG policy. Both systems gave users instructions to set up their Internet connection

Type : Audio and text

Recording conditions : Microphone.

Experimental setup : Wizard of Oz.

Type of annotation : Manual.

Annotation scheme : User's utterances were annotated but not transcribed

Size : Data Collection - 272MB (18 dialogues). Evaluation - 516 MB (36 dialogues)

Original aim of data : Learning User modelling strategies for NLG optimization

Date of collection : February 2009 - April 2010

2.8 FT Appointment Scheduling Systems 3 and 4

Partner in charge : FT

Description : This corpus contains all the raw traces collected by the FT platform during the final CLASSiC evaluation. It contains the following directories:

- `audio` contains all the user acoustic signal;
- `disserto` contains the platform logs for the frontend, System 3 and 4
- `svip` contains the platform logs for the vocal platform

In addition, the `socio` directory contains a selection of dialogues from Systme 3 and 4 whose audio files were processed to also include the system utterances. This selection was created for manual analysis. It contains 6 subdirectories :

- best : a selection of calls from the 10 users which produced the most efficient dialogues;
- mismatch-SKO-UOK : a selection of calls from 5 users erroneously thinking they have got a right appointment;
- mismatch-SOK-UKO : a selection of calls from 5 users thinking they have got no appointment, whereas the system has booked one;
- persevere-fail : a selection of calls from 5 users trying very hard to get an appointment, but finally failing;
- persevere-OK : a selection of calls from 5 users trying very hard to get an appointment, with a final success;
- recovery : a selection of calls from 10 users when the system has trouble understanding them.

Type : Audio and text

Recording conditions : Telephone.

Experimental setup : Real System.

Type of annotation : Automatic.

Annotation scheme : The raw files are completed with some description and summary files. The intermediate summary is presented in `classic.complete.csv`. This file is a comma-separated-value text file, with a header. It contains the following fields:

- code: the unique dialogue code;
- svipfile: for systems 3 and 4, the log file for the vocal platform;
- conid: for systems 3 and 4, the connection id on the vocal platform;
- frontendfile: the logfile for the frontend part of the call;
- service: the service called (2,3 or 4);
- servicefile: for system 3 and 4, the log file for the DM;

- serviceId: for system 3 and 4, the unique id for the DM;
- user: the user calendar;
- system: for system 2, the phoneline called (there's a different phoneline for each system scenario);
- audiofile: the client-side audio file for the dialogue;
- result: the task completion as perceived by the system;
- appointment: the last timeslot given by the system;
- CORRECT: whether the appointment timeslot is valid according to the user's agenda;
- time: to be ignored (initially thought for the call time and date);
- user_difficulty: the scenario's difficulty for the user;
- system_difficulty: the scenario's difficulty for the system;
- appointment_thought_correct_by_the_user: whether the user thinks she got a valid appointment;
- asr_score: asr score (on a 1-6 level);
- nlg_score: nlg score (on a 1-6 level);
- tts_score: tts score (on a 1-6 level);
- future_use: whether the user would reuse such a system (on a 1-6 level);
- global_score: a global appreciation score (on a 1-10 level);
- nb_rejects: for systems 3 and 4, number of asr rejections;
- duration: for systems 3 and 4, call duration;

Given all the functional problems encountered during the evaluation, we have decided to filter this intermediate file to extract the dialogues corresponding to the system "nominal conditions" (*i.e.* when the system was not down due to a network or licence problem). The list of all dialogues kept are in `valid_ids.txt` . It was used to produce the final evaluation basis `classic.filtered.csv`, which follows the same format as `classic.complete.csv`.

The `System2.csv`, `System3.csv` and `System4.csv` contain a summary of the user questionnaire fields, except for the commentary fields, which cannot be automatically processed.

Size : 10 GB

Original aim of data : Evaluation of Appointment Scheduling systems

Date of collection : November 8^{th} 2010 to February 3^{rd} 2011.

2.9 1013 corpus

Partner in charge : FT

Description : The 1013 corpus contains raw system logs and audio files for both the lab version (`lab` directory) and the commercially deployed (`prod` directory) 1013+ Appointment Scheduling System.

Directories description :

- `raw` contains the raw dialogues in a daily zip file. They contains both a text log file and all the acoustic signal processed by the speech recognizer;
- `transcribed` contains the manually annotated dialogues in a daily textfile;
- `dialogues` contains the reconstituted dialogues in a text format conforming to the dialogue acts described in deliverables D5.1.2 and D6.1.3.

The `script` directory contains the PERL script used to reconstitute the `dialogues` files from the `transcribed` files.

Type : Audio and text

Recording conditions : Telephone.

Experimental setup : Real System.

Type of annotation : Automatic.

Annotation scheme : See Deliverable D5.1.2

Size : It represents 10000 dialogues

Original aim of data : Evaluation of the 1013+ system

Date of collection : December 16^{th} 2009 to January 12^{th} 2011.

2.10 TownInfo - NLG WoZ Corpus

Partner in charge : UEDIN

Description : A Wizard-of-Oz (WoZ) corpus to study Information Presentation strategies for Spoken Dialogue Systems. We designed and implemented a server-client, web-based WoZ tool to have collected the data to study different Information Presentation (IP) strategies as explored by human 'wizards'.

In particular, we investigate which IP actions the wizard selects in a specific dialogue context, given uncertainty about the user's input, varying number of database matches, stochastic surface realisation, and varying user behaviour.

We used this data to optimise IP strategies for Spoken Dialogue Systems (SDS) using Reinforcement Learning (RL), where we follow a new framework of NLG as planning under uncertainty [4], [8],[5].

The WoZ experiments were reported in detail at [6],[9],[10],[1].

Type : Audio and text.

Recording conditions : Headset microphone for Skype running on a PC or a normal telephone via an Edinburgh local number (0131)2085287.

The user voice was recorded by PowerGramo[4] software. The Wizard voice is synthesized by Cereproc[5] TTS.

The audio is transmitted and recorded at the server side.

Experimental setup : The experiments were done by using the WoZ tool developed by us. The tool is a server-client, web-based Java application developed on Netbeans (V6.1) using Visual Web JavaServer Faces framework and Apache Tomcat web server (V6.0.16).

There were three types of participants in the experiments: the user, the wizard and the experimenter. The wizard was sitting in the room of the server side. The user and the experimenter were sitting in the room of the user side. All three participants had a computer running the WoZ tool with different interfaces to communicate each other. In the non noise model the experimenter would do nothing except for giving initial instructions to the user, while in the noise model he/she would also use the noise model interface to communicate with the wizard.

The WoZ tool therefore also includes a noise simulation and a Natural Language Generation (NLG) surface realiser.

The user tasks in the experiments were to book restaurants in Edinburgh where we used a real database of Edinburgh restaurants provided by TheList[6].

Type of annotation : Automatic semantic annotation by loggings.

Annotation scheme : Automatic semantic annotation using an extended version of CLASSiC data schema – D3.1: Shared Context Model (XML Schema).

Size : 1.6MB zip file for an anonymised version of the data (xml and feature files), 2.2GB audio files.

We have collected 213 dialogues with 18 subjects and 2 wizards in this setup. Each user has to perform a total of 12 tasks, where no task set is seen twice by one wizard. After each task the user answers a questionnaire on a 6 point Likert scale. The data contains ca. 2236 utterances in total: 1465 system prompts and ca. 771 user prompts. We also automatically extracted 81 features from the XML logfiles.

The data analysis results are reported in the Deliverable D4.2[10].

Original aim of data : Following the framework of NLG as 'planning under uncertainty'[5] to investigate the wizards' Information Presentation (IP) decisions in the presence of uncertainty. Using statistical machine learning techniques to learn the optimal NLG IP strategies for Spoken Dialogue Systems from the data.

Date of collection : Feb.- March 2009

[4] http://www.powergramo.com/
[5] http://www.list.co.uk/
[6] http://www.list.co.uk/

Bibliography

[1] C. Boidin, V. Rieser, S. Janarthanam, and O. Lemon. Domain-limited TTS corpus for expressive speech synthesis and Wizard-of-Oz Data for NLG Strategies. Technical Report D6.1.1, CLASSiC Deliverable, 2009.

[2] S. Janarthanam, H. Hastie, O. Lemon, and X. Liu. 'The day after the day after tomorrow? ' A machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In *Proceedings of SIGDIAL*, 2011. (to appear).

[3] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*, pages 79–86, Phuket, Thailand, 2005.

[4] O. Lemon. Adaptive Natural Language Generation in Dialogue using Reinforcement Learning. In *Proceedings of SEMdial*, 2008.

[5] O. Lemon. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Proceedings of EACL*, 2009.

[6] X. Liu, V. Rieser, and O. Lemon. A Wizard-of-Oz interface to study Information Presentation strategies for Spoken Dialogue Systems. In *Proceedings of the 1st International Workshop on Spoken Dialogue Systems*, 2009.

[7] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:31:71–105, 2005.

[8] V. Rieser and O. Lemon. Learning human multimodal dialogue strategies. *Journal of Natural Language Engineering*, 2009.

[9] V. Rieser, O. Lemon, and X. Liu. Optimising information presentation for spoken dialogue systems. In *Proceedings of ACL 2010*, 2010.

[10] V. Rieser, X. Liu, and O. Lemon. Optimal Wizard NLG Behaviours in Context. Technical Report D4.2, CLASSiC Deliverable, 2009.

[11] M. Surdeanu, R. Johansson, A. Meyers, L. Màrguez, and J. Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL-2008*, Manchester, UK, 2008.